Epidemics on contact networks: a general stochastic approach

Pierre-André Noël, Antoine Allard, Laurent Hébert-Dufresne, Vincent Marceau and Louis J. Dubé
Département de Physique, de Génie Physique et d'Optique, Université Laval, Québec (QC), Canada

December 8, 2011

Abstract

Dynamics on networks is considered from the perspective of Markov stochastic processes. We partially describe the state of the system through network motifs and infer any missing data using the available information. This versatile approach is especially well adapted for modelling spreading processes and/or population dynamics. In particular, the generality of our systematic framework and the fact that its assumptions are explicitly stated suggests that it could be used as a common ground for comparing existing epidemics models too complex for direct comparison, such as agent-based computer simulations. We provide many examples for the special cases of susceptible-infectious-susceptible (SIS) and susceptible-infectious-removed (SIR) dynamics (*e.g.*, epidemics propagation) and we observe multiple situations where accurate results may be obtained at low computational cost. Our perspective reveals a subtle balance between the complex requirements of a realistic model and its basic assumptions. Keywords: contact networks, epidemics, stochastic processes, complex networks, spreading dynamics, Markov processes

1 Introduction

Mathematical modelling has proven a valuable tool when addressing public health issues. The increase in availability of powerful computer resources has facilitated the use of agent-based models and other complex modelling approaches, all accounting for numerous parameters and assumptions [1, 2, 3]. Our confidence in these models may increase when they are shown to agree with empirical observations and/or with previously accepted models. However, when discrepancies appear, the complexity of these computer programs may obfuscate the effect of underlying assumptions, making it difficult to isolate the source of disagreement. While analytical approaches offer more insights on the underlying assumptions, their use is often restricted to simpler interaction structures and/or dynamics.

The purpose of this paper is to systematically model the global behaviour of stochastic systems composed of numerous elements interacting in a complex way. "Complex" here implies that interactions among the elements follow some nontrivial patterns that are neither perfectly regular nor completely random, as often seen in real-world systems. "Stochastic" implies that the system may not be completely predictable to us and that a probabilistic solution is sought.

To this end, we present (Sec. 2) a general modelling scheme where network theory [4, 5] accounts for the interactions between the elements of the system and where a birth-death Markov process [6] models the stochastic dynamics. Since a tremendous amount of information may be required to store the state of the whole system, we seek the part of this information that is important for the problem at hand and then approximate the dynamics by tracking only this limited subset. Part of the discarded data may still affect, albeit weakly, the behaviour of the system. We fill this knowledge gap by inferring the missing information such that it is consistent both with the information we follow and any other prior information that is available to us.

An important part of this paper (Sec. 3) provides explicit examples to these general ideas. For simplicity, each case either corresponds to a susceptible-infectious-susceptible (SIS) or to a susceptible-infectious-removed (SIR) spreading processes, both standards in the study of infectious diseases propagation. While our first examples study simpler cases, facilitating the understanding of our systematic method, the later models show how the same approach applies to more complex interaction structures.

We then compare and analyse the results of these examples (Sec. 4). This reveals some general considerations for both the accuracy and the complexity of our modelling approach. We find that treating the inferences of missing information explicitly helps systematize the model development and highlights numerous possibilities for future developments. An

important simplification occurs for SIR spreading processes and related dynamics, leading to an *exact* model with a small number of dynamical variables.

We conclude (Sec. 5) on how our general approach may be applied beyond spreading processes, for example, population dynamics. Returning to the problem of understanding the source of discrepancies in complex models, we explain how *modelling these models* with our method could help identifying important assumptions and isolating the source of disagreement. Mathematical details and further generalizations are also presented in an Electronic Supplementary Material (ESM) [7].

2 General modelling scheme

We assume that the real-world system to be modelled is sufficiently well understood to implement a Monte Carlo computer simulation that approximately reproduces its behaviour. We refer to this hypothetical computer simulation as the *full system*: Z denotes the *state* of the full system (*i.e.*, all the data that would be stored by the computer program) while V denotes the *rules* governing the evolution of Z in time (*i.e.*, the program itself).

However, there are many situations where a direct implementation of the full system is impractical due to storage and/or computation considerations. We thus design a *simplified system* that aims at reproducing the behaviour of the full one, while requiring less resources.

The state X of this simplified system (*much* smaller than Z) evolves in time according to the rules W. Moreover, we note Y any known *prior information* that is relevant in a Bayesian inference of Z

$$P(Z|X,Y) = \frac{P(Z|Y)\,P(X|Y,Z)}{P(X|Y)} \quad . \tag{1}$$

This last point is crucial: Y often makes the difference between an accurate model and a useless one. It bridges the gap between the simplified representation of the state of the system (*i.e.*, X) and the full one (*i.e.*, Z).

Since we are interested in systems composed of many elements interacting through complex patterns, we express the previous quantities in terms of networks.

2.1 Networks

A *network* (graph) is a collection of *nodes* (vertices) and *links* (edges). Nodes model the elements of a system; links join nodes pairwise to represent interactions between the corresponding elements. Two nodes sharing a link are said to be *neighbours* and the *degree* of a node is its number of neighbours. The part of a link that is attached to a node is called a *stub*: there are two stubs per link and each node is attached to a number of stubs equal to its degree. A link with both ends leading to the same node is called a *self-loop* and *repeated links* occur when more than one link join the same two nodes.

There may be systems such that specifying its state Z exactly amounts to specifying the network structure. However, most systems are not purely structural: they are composed of elements that, by themselves, require additional information to be properly characterized. Hence, we assign to each node a *node state* that specifies the intrinsic properties of the corresponding element in the system. Both the structure and these intrinsic node states are specified by Z; see ESM [7] §I for further examples of information that may be contained in Z, including the important case of directed networks.

2.2 Motifs

Specifying the complete structure of a complex network requires a tremendous amount of information. Since we want the state X of a simplified system to be of manageable size, approximations have to be made. A convenient way to do so, and one that has proven to give good results in the past [8, 9, 10, 11, 12, 13, 14], is to specify the network structure through its building blocks.

A network *motif* is a pattern that may appear a number of times in the network. For example, two linked nodes form a *pair motif* while three nodes all neighbours of one another form a *triangle motif*. Motifs may encode intrinsic node states or other relevant information; further details and examples are provided throughout Section 3 as well as in ESM [7] §II.

We define the *state vector* $\mathbf{x}$ of a system as a vector of integers specifying how many times different motifs appear in the network. These motifs may be attached together to form a network structure: the state vector $X = \mathbf{x}$ enumerates the

available building blocks while the prior information Y specifies how such blocks may be attached. There will usually be numerous valid ways to attach the blocks, some more probable than others. Given the available information, the resulting distribution is our best estimate for $P(Z|\mathbf{x}, Y)$.

By judiciously choosing the motifs enumerated in $\mathbf{x}$ and by specifying informative prior information Y, one may hope for this probability distribution to be densely localized around the "real" value of Z in the full system. This mapping can then be used to convert the rules V of the full system to the rules W of the new simplified one. We approach this problem from the perspective of birth-death Markov processes.

2.3 Birth-death stochastic processes

In a *birth-death* process, the elements composing a system may be destroyed (death) while new ones may be created (birth). It is therefore natural to state the rules W of our simplified system in those terms: any change in the state vector $\mathbf{x}$ may be perceived as an event where motifs are created and/or destroyed.

Quantitatively, a *forward transition event of type* j takes the system from state $\mathbf{x}$ to state $\mathbf{x} + \mathbf{r}^j$ and has probability $q_j^+(\mathbf{x}, Y)\,dt$ to occur during the time interval $[t, t + dt)$. Similarly, a *backward transition event of type* j takes the system from state $\mathbf{x}$ to state $\mathbf{x} - \mathbf{r}^j$ and has probability $q_j^-(\mathbf{x}, Y)\,dt$ to occur during the same time interval.

Specifying for each j the elements r_i^j of the *shift vector* $\mathbf{r}^j$ together with the *rate functions* $q_j^+(\mathbf{x}, Y)$ and $q_j^-(\mathbf{x}, Y)$ thus completely define the rules W governing the simplified system. This Markov process is summarized in the master equation

$$\frac{dP(\mathbf{x}|Y,t)}{dt} = \sum_j \Big[q_j^+\big(\mathbf{x} - \mathbf{r}^j, Y\big)\, P\big(\mathbf{x} - \mathbf{r}^j \big| Y, t\big) - q_j^+(\mathbf{x}, Y)\, P(\mathbf{x}|Y,t) + q_j^-\big(\mathbf{x} + \mathbf{r}^j, Y\big)\, P\big(\mathbf{x} + \mathbf{r}^j \big| Y, t\big) - q_j^-(\mathbf{x}, Y)\, P(\mathbf{x}|Y,t) \Big] \tag{2}$$

specifying the evolution of the probability $P(\mathbf{x}|Y,t)$ to observe state $\mathbf{x}$ at time t (notation compatible with [6] §7.5).

We now consider two approximations that are often justified for large systems: the elements of $\mathbf{x}$ may be treated as varying continuously and the probability distribution is strongly concentrated around its mean value. In such cases, the evolution of the mean value $\boldsymbol{\mu}(t) = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x}|Y,t)$ for the vector $\mathbf{x}$ at time t is approximately given by

$$\frac{d\boldsymbol{\mu}(t)}{dt} = \mathbf{a}\big(\boldsymbol{\mu}(t)\big) \tag{3a}$$

$$a_i(\mathbf{x}) = \sum_j r_i^j \Big[q_j^+(\mathbf{x}, Y) - q_j^-(\mathbf{x}, Y) \Big] \tag{3b}$$

where we defined the *drift vector* $\mathbf{a}(\mathbf{x})$ of elements $a_i(\mathbf{x})$ (see [6] §7.5.3 and §4.4.9).

In order to further refine our knowledge of $P(\mathbf{x}|Y,t)$ in the vicinity of this deterministic solution, we define the *evolution matrix* $A(t,t')$, the *diffusion matrix* $B(\mathbf{x})$ $\big($of elements $B_{ii'}(\mathbf{x})\big)$and the *covariance matrix* $C(t)$

$$A(t,t') = \exp\left[\int_{t'}^{t} J_{\mathbf{a}}\big(\boldsymbol{\mu}(t'')\big) dt'' \right] \tag{4a}$$

$$B_{ii'}(\mathbf{x}) = \sum_j r_i^j r_{i'}^j \Big[q_j^+(\mathbf{x}, Y) + q_j^-(\mathbf{x}, Y) \Big] \tag{4b}$$

$$C(t) = \int_0^t A(t,t') \cdot B\big(\boldsymbol{\mu}(t')\big) \cdot A(t,t')^T dt' \tag{4c}$$

where $J_{\mathbf{a}}(\mathbf{x})$ is the Jacobian matrix of $\mathbf{a}$ evaluated at $\mathbf{x}$. Noting d the size of the vector $\mathbf{x}$, the probability distribution may be approximated by a d-dimensional Gaussian

$$P(\mathbf{x}|Y,t) = \frac{\exp\left\{ -\frac{1}{2} \big[\mathbf{x}(t) - \boldsymbol{\mu}(t)\big]^T \cdot C(t)^{-1} \cdot \big[\mathbf{x}(t) - \boldsymbol{\mu}(t)\big] \right\}}{\sqrt{(2\pi)^d \big|C(t)\big|}} \tag{5}$$

where $\left|C(t)\right|$ is the determinant of $C(t)$. Note that (2)–(5) are all textbook relationships.

Although many other tools are available for the analysis of stochastic systems, the simplicity, the generality and the straightforwardness of the Gaussian approximation make it an instrument of choice that will be used extensively in this article.

3 Application to spreading dynamics

Without prejudice to the generality of Section 2, we now focus our study to spreading processes [15, 16, 17]. An epidemiological terminology is used: whatever propagates among neighbouring nodes, be it desirable or not, is called an *infection*. We find that the basic SIS and SIR epidemiological models, both to be defined shortly, require little prior knowledge from the part of the reader while being sufficiently complex for the needs of the present study.

At a given time, the intrinsic state of each node of an *SIS model* may either be *Susceptible* (not carrying the infection) or *Infectious* (carrying the infection). The full system state Z hence specifies each node's intrinsic state together with the complete structure of the network. The rules V are simple: during any time interval $[t, t + dt)$, each infectious node may recover (*i.e.*, it becomes susceptible) with probability $\alpha\, dt$ and, for each of its susceptible neighbours, has probability $\beta\, dt$ to transmit the infection (*i.e.*, the neighbour becomes infectious).

In addition to the susceptible and infectious intrinsic states, the nodes of an *SIR model* may also be *Removed* (once had the infection and can neither acquire nor transmit it ever again). The rules V are the same than for the SIS model with respect to infection (*i.e.*, infectious nodes transmit to their susceptible neighbour with probability $\beta\, dt$), but recovery is replaced by removal (*i.e.*, infectious nodes become removed with probability $\alpha\, dt$).

The remainder of this section studies how different choices of state vector $\mathbf{x}$ and prior information Y translate in the rules W of the simplified system. In each case, W is defined through a set of equations whose tags all share the same numeral, *e.g.*, (6a)–(6f). Although figures present results concomitantly with the specification of the corresponding models, all discussions are delayed to Section 4.

3.1 Pair-based SIS model

Section 2.2 defined a pair motif as two linked nodes. Since the nodes of a SIS model are either susceptible or infectious, there are three possibilities for pair motifs: two linked susceptible nodes (noted S–S), two linked infectious nodes (noted I–I) and a susceptible node linked to an infectious one (noted S–I). Two nodes involved in a pair motif may have other neighbours.

Pair motifs are often used in conjunction with *node motifs*: the trivial structure that is one node. In the SIS model, there are two possibilities for a node motif: susceptible nodes (noted S) and infectious nodes (noted I). A state vector $\mathbf{x}$ based on both node and pair motifs would thus be composed of five elements enumerating the amount of times each motif appears in the network: x_S, x_I, $x_{S\text{–}S}$, $x_{S\text{–}I}$ and $x_{I\text{–}I}$. However, additional assumptions about the structure of the network may cause some of these quantities to be redundant.

3.1.1 Degree-regular network

We first consider the simple case where the network is known to be a *n-regular network* of size N: there are N nodes in the network which all have n neighbours (degree n). Such a network must respect the structural constraints $x_S = N - x_I$, $x_{S\text{–}S} = \frac{1}{2}\left(nx_S - x_{S\text{–}I}\right)$ and $x_{I\text{–}I} = \frac{1}{2}\left(nx_I - x_{S\text{–}I}\right)$. Hence, with the prior information Y specifying N and n, the state vector

$$\mathbf{x} = \left(x_I, x_{S\text{–}I}\right) \tag{6a}$$

suffices to obtain all the five node and pair motifs.

In those terms, the rules V specify that an infection has probability $\beta\, x_{S\text{–}I}\, dt$ to occur during the time interval $[t, t+dt)$ while a recovery has probability $\alpha\, x_I\, dt$ to occur. Clearly, an infection translates to the destruction of a S motif and the creation of a new I one, and a recovery corresponds to the inverse process. However, pair motifs are also affected by such transitions since the affected node had neighbours. Hence, the effect on $\mathbf{x}$ of the infection or recovery of a node depends on some information that is not directly tracked by $\mathbf{x}$ — *i.e.*, what is the state of the infected or recovered node's neighbours — and we thus have to *infer* this information from the available data.

In order to facilitate this inference, we define the *first neighbourhood motif* $S\Gamma_1(k_S, k_I)$ as a susceptible node that has k_S susceptible neighbours and k_I infectious neighbours. Similarly, the motif $I\Gamma_1(k_S, k_I)$ corresponds to an infectious node with k_S susceptible neighbours and k_I infectious ones. In both cases, we qualify as *central* the node whose neighbours are explicitly stated. The other nodes of the first neighbourhood motif, *i.e.*, the neighbours of the central node, may have other neighbours of their own.

We can now define a forward transition event of type $j \in \{0, 1, \ldots, n\}$ as the infection of the central node of a $S\Gamma_1(n-j, j)$ motif. In terms of node and pair motifs, this implies the destruction of one of the S motifs, of $n-j$ of the $x_{S\text{–}S}$ motifs and of j of the $x_{S\text{–}I}$ motifs together with the creation of one new I motif, of $n-j$ new $x_{S\text{–}I}$ motifs and of j new $x_{I\text{–}I}$ motifs. Since only x_I and $x_{S\text{–}I}$ are tracked, the shift vectors are

$$\mathbf{r}^j = \big(1, n-2j\big) \quad . \tag{6b}$$

This same vector also defines the backward transition events $j \in \{0, 1, \ldots, n\}$ which correspond to the recovery of the central node of a $I\Gamma_1(n-j, j)$ motif.

Looking back at the rules V, the corresponding forward and backward transition rate functions are

$$q_j^+(\mathbf{x}, Y) = \beta\, x_{S\text{–}I}\, P(S\Gamma_1(n-j, j) | \, S, j \geq 1, \mathbf{x}, Y) \tag{6c}$$

$$q_j^-(\mathbf{x}, Y) = \alpha\, x_I\, P(I\Gamma_1(n-j, j) | \, I, \mathbf{x}, Y) \tag{6d}$$

where two inference terms have been defined.

The inference term of (6d) gives the probability for a motif to be a $I\Gamma_1(n-j, j)$ knowing that it has an infectious node at its center and that the current state vector is $\mathbf{x}$ with prior information Y. For a sufficiently large network, this approximately corresponds to randomly drawing the n neighbours of the central infectious node among the pair motifs S–I and I–I

$$P(I\Gamma_1(n-j, j) | \, I, \mathbf{x}, Y) = \binom{n}{j} \left(\frac{x_{S\text{–}I}}{n\, x_I}\right)^j \left(1 - \frac{x_{S\text{–}I}}{n\, x_I}\right)^{n-j} \quad . \tag{6e}$$

The inference term of (6c) is very similar except that the central susceptible node is known to have at least one infectious neighbours since it acquired the infection through a S–I motif

$$P(S\Gamma_1(n-j, j) | \, S, j \geq 1, \mathbf{x}, Y) = \binom{n-1}{j} \left(\frac{x_{S\text{–}I}}{n\,(N - x_I)}\right)^j \left(1 - \frac{x_{S\text{–}I}}{n\,(N - x_I)}\right)^{n-1-j} \tag{6f}$$

which complete the rules W for the pair-based SIS model on a n-regular network of size N.

Figure 2 compares the results produced by this simplified model (defined by W, $\mathbf{x}$ and Y) to the corresponding full one (defined by V and Z). Figure 3 shows the probability distribution for the same data. Note that, although presented differently, this model corresponds to the one presented in [18]; Fig. 1 is provided for comparison with Fig. 2(c) of [18].

3.1.2 Erdős-Rényi network

We now consider the case where the network is an Erdős-Rényi network: there are N nodes in the networks and M links are randomly assigned. This knowledge constrains two of the five node and pair motifs (*i.e.*, $x_S = N - x_I$ and $x_{S\text{–}S} = M - x_{S\text{–}I} - x_{I\text{–}I}$) and a state vector of three elements suffices

$$\mathbf{x} = \big(x_I, x_{S\text{–}I}, x_{I\text{–}I}\big) \quad . \tag{7a}$$

The method used in Section 3.1.1 has to be adapted since the degree of each node is not constrained to a single value. Indeed, a susceptible node that gets infected may *a priori* be the center of any of the $S\Gamma_1(k_S, k_I)$ motifs. Still, we could design a bijective mapping between the vector of integers $\mathbf{k} = (k_S, k_I)$ and an event type j.

The details of the chosen mapping do not matter: we simply define the forward transition event of type $\mathbf{k}$ as the infection of the central node of a $S\Gamma_1(k_S, k_I)$ motif, which may conveniently be noted $S\Gamma_1(\mathbf{k})$ instead. Similarly, the

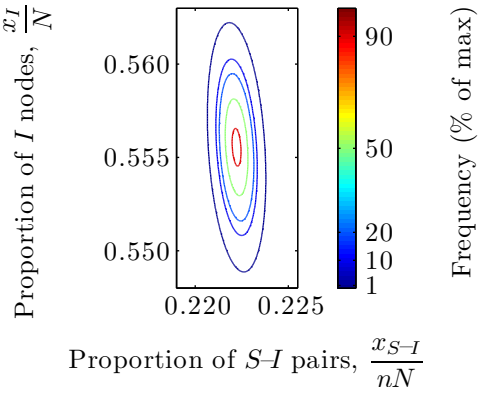


Figure 1: (Online version in colour.) Distribution of post-transient ($t \to \infty$) outcomes as predicted by the SIS model on regular network (6) using the approximations (3)–(4). The axes (proportion of I among node motifs *vs* proportion of S–I among pair motifs), network structure ($N = 10^5$ nodes, each of degree $n = 5$) and parameters ($\alpha = 0.1^2$ and $\beta = 0.05$) are the same as for Fig. 2(c) from [18]. Frequencies (in percent) are used to facilitate comparison with [18]; they are simply obtained from the probability densities of (5) multiplied by 100.

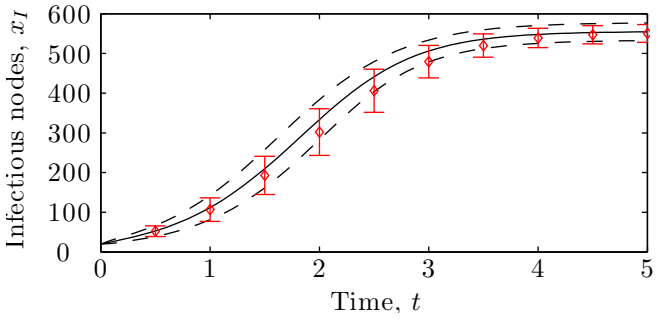


Figure 2: (Online version in colour.) Time evolution of the number of infectious nodes x_I for SIS dynamics ($\alpha = 2$ and $\beta = 1$) on a regular network of $N = 10^3$ nodes (20 initially infectious) of degree $n = 5$. Curves: results for the simplified system (6) approximated with (3)–(4). The continuous curve shows the mean value while the dashed curves delimit the range of one standard deviation above and below the mean. Symbols: averaged results of 10^5 numerical simulations of the full system. The parameters α and β correspond to those of Fig. 1 after rescaling the time unit.

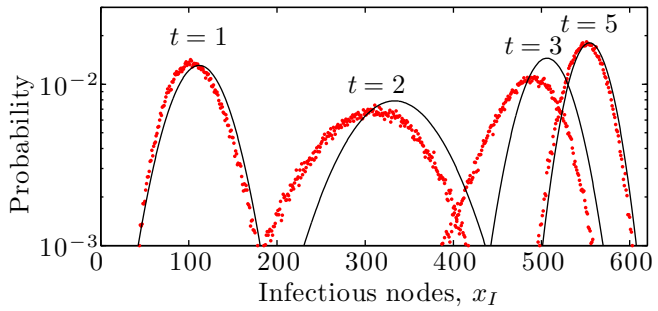


Figure 3: (Online version in colour.) Probability distribution at different times for the number of infectious nodes x_I. All parameters are the same as in Fig. 2. Curves: Gaussian approximation for the simplified system. Symbols: binned results of 10^5 numerical simulations of the full system.

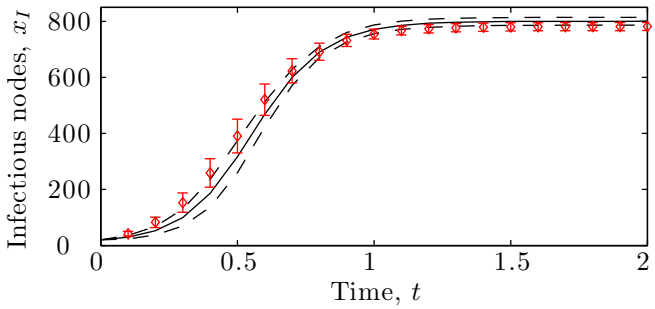


Figure 4: (Online version in colour.) Time evolution of the number of infectious nodes x_I for SIS dynamics ($\alpha = 2$ and $\beta = 1$) on an Erdős-Rényi network of $N = 10^3$ nodes (20 initially infectious) and $M = 5 \cdot 10^3$ links. The mean and range of one standard deviation above and bellow the mean are shown. Curves: simplified system. Symbols: full system (10^5 simulations).

backward transition event of type $\mathbf{k}$ is defined as the recovery of the central node to a $I\Gamma_1(\mathbf{k})$ motif. The corresponding shift vector and rate functions are

$$\mathbf{r}^{\mathbf{k}} = \big(1, k_S - k_I, k_I\big) \tag{7b}$$

$$q^{+}_{\mathbf{k}}(\mathbf{x}, Y) = \beta\, x_{S\text{–}I}\, P(S\Gamma_1(\mathbf{k}) |\, S, k_I \geq 1, \mathbf{x}, Y) \tag{7c}$$

$$q^{-}_{\mathbf{k}}(\mathbf{x}, Y) = \alpha\, x_I\, P(I\Gamma_1(\mathbf{k}) |\, I, \mathbf{x}, Y) \tag{7d}$$

where the inference terms bear the same meaning as their previous counterpart. Again assuming large network size, these inference terms are obtained by evaluating the probability for each pair motif to include the considered node. Hence, the products of binomial distributions

$$P(S\Gamma_1(\mathbf{k}) |\, S, l \geq 1, \mathbf{x}, Y) = \binom{2\,x_{S\text{–}S}}{k_S} x_S^{-k_S} \big(1 - x_S^{-1}\big)^{2\,x_{S\text{–}S} - k_S} \binom{x_{S\text{–}I} - 1}{k_I - 1} x_S^{-(k_I - 1)} \big(1 - x_S^{-1}\big)^{x_{S\text{–}I} - k_I} \tag{7e}$$

$$P(I\Gamma_1(\mathbf{k}) |\, I, \mathbf{x}, Y) = \binom{x_{S\text{–}I}}{k_S} x_I^{-k_S} \big(1 - x_I^{-1}\big)^{x_{S\text{–}I} - k_S} \binom{2\,x_{I\text{–}I}}{k_I} x_I^{-k_I} \big(1 - x_I^{-1}\big)^{2\,x_{I\text{–}I} - k_I} \tag{7f}$$

complete the rules W for the pair-based SIS model on a Erdős-Rényi network of size N with M links.

Figure 4 compares the results produced by this simplified model to the corresponding full one. Section 4.2 discusses these results and provides further details concerning pair-based models.

3.2 First neighbourhood SIS model

We consider a full model (V and Z) for SIS dynamics on a *configuration model* (CM) network: given a sequence $\{n_0, n_1, n_2, \cdots\}$, links are randomly assigned between nodes such that, for each degree κ, there are n_κ nodes of degree κ. In a computer simulation, we create n_κ nodes with κ stubs for each possible κ and then randomly pair stubs to form links. No particular mechanism is used to prevent the formation of repeated links and self-loops: this simplifies the analytical treatment and has little effect when the network size is sufficiently large.

Our simplified model handles the heterogeneity in node degree by enumerating every possible first neighbourhood motifs in its state vector

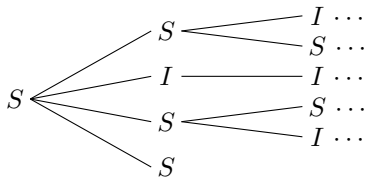
$$\mathbf{x} = \big(x_{S\Gamma_1(0,0)}, x_{S\Gamma_1(1,0)}, \cdots, x_{I\Gamma_1(0,0)}, \cdots\big) \quad . \tag{8a}$$

Although this vector should be infinite in the general case, it is not the case when, *e.g.*, the prior information Y states that no node has a degree superior to $\mathcal{K}$.

For the same reasons that models tracking node and pair motifs (Section 3.1) had their transition events defined in terms of first neighbourhood motifs, the transition events are here defined in terms of *second neighbourhood motifs*: a central node, its neighbours and the neighbours of those neighbours. In the same way that we note $\nu\Gamma_1(\mathbf{k})$ the first

neighbourhood motif formed by a state ν central node with neighbourhood specified by $\mathbf{k}$, we note $\nu\Gamma_2(\mathbf{K})$ the second neighbourhood motif formed by a state ν central node with neighbourhood specified by $\mathbf{K}$.

The elements of $\mathbf{K}$ may be indexed with first neighbourhood motifs: the central node has $K_{\nu'\Gamma_1(\mathbf{k}')}$ state ν' neighbours whose other neighbours (*i.e.*, excluding the central node) are specified by $\mathbf{k}'$. Hence, the second neighbourhood motif



is noted $S\Gamma_2(\mathbf{K})$ with all elements of $\mathbf{K}$ zero except for $K_{S\Gamma_1(0,0)} = 1$, $K_{I\Gamma_1(0,1)} = 1$ and $K_{S\Gamma_1(1,1)} = 2$. Note that the central node of this second neighbourhood motif is also the central node of the first neighbourhood motif $S\Gamma_1(3,1)$. In general, we note $\nu\widetilde{\Gamma}_1(\mathbf{K})$ the first neighbourhood motif that shares the same central node as the second neighbourhood motif $\nu\Gamma_2(\mathbf{K})$.

We digress further to introduce the *unit vector* notation $\widehat{\mathbf{e}}_{\mathcal{M}}$ where $\mathcal{M}$ represents a motif; all the elements of this vector are zero except for the $\mathcal{M}$-th, which is one. The total number of elements in $\widehat{\mathbf{e}}_{\mathcal{M}}$ should be clear from the context. As a concrete example, the right hand side of (6b) could be noted $\widehat{\mathbf{e}}_I + (n-2j)\widehat{\mathbf{e}}_{S\text{–}I}$.

Similarly to Section 3.1.2, we define the forward transition event of type $\mathbf{K}$ to be the infection of the central node of a $S\Gamma_2(\mathbf{K})$ motif and the backward transition event of type $\mathbf{K}$ as the recovery of the central node of a $I\Gamma_2(\mathbf{K})$ motif. The corresponding shift vector is

$$\mathbf{r}^{\mathbf{K}} = \widehat{\mathbf{e}}_{I\widetilde{\Gamma}_1(\mathbf{K})} - \widehat{\mathbf{e}}_{S\widetilde{\Gamma}_1(\mathbf{K})} + \sum_{\nu}\sum_{\mathbf{k}} K_{\nu\Gamma_1(\mathbf{k})}\big(\widehat{\mathbf{e}}_{\nu\Gamma_1(\mathbf{k}+\widehat{\mathbf{e}}_I)} - \widehat{\mathbf{e}}_{\nu\Gamma_1(\mathbf{k}+\widehat{\mathbf{e}}_S)}\big) \quad . \tag{8b}$$

The first line shows the direct effect of a change of state in the central node while the second one handles the "collateral effect" on its immediate neighbours. Here the unit vector $\widehat{\mathbf{e}}_\nu$ has the same dimension as $\mathbf{k}$ (*i.e.*, two) while $\widehat{\mathbf{e}}_{\nu\Gamma_1(\mathbf{k})}$ has the same dimension as $\mathbf{x}$. Sums are taken over all the accessible values of ν and $\mathbf{k}$.

The corresponding rate functions are

$$q^+_{\mathbf{K}}(\mathbf{x},Y) = \beta x_{S\widetilde{\Gamma}_1(\mathbf{K})}\Big(\sum_{\mathbf{k}} K_{I\Gamma_1(\mathbf{k})}\Big) P\Big(S\Gamma_2(\mathbf{K})\Big|\, S\widetilde{\Gamma}_1(\mathbf{K}),\mathbf{x},Y\Big) \tag{8c}$$

$$q^-_{\mathbf{K}}(\mathbf{x},Y) = \alpha\, x_{I\widetilde{\Gamma}_1(\mathbf{K})}\, P\Big(I\Gamma_2(\mathbf{K})\Big|\, I\widetilde{\Gamma}_1(\mathbf{K}),\mathbf{x},Y\Big) \quad . \tag{8d}$$

Note that, unlike (6c)–(6d) and (7c)–(7d), the inference terms in (8c)–(8d) have the same form: the probability for a motif to be a $\nu\Gamma_2(\mathbf{K})$ knowing (in addition to $\mathbf{x}$ and Y) that its central node is also the central node of a $\nu\widetilde{\Gamma}_1(\mathbf{K})$ motif. Again assuming a large network size, they are provided by a product of multinomial distributions

$$P\Big(\nu\Gamma_2(\mathbf{K})\Big|\,\nu\widetilde{\Gamma}_1(\mathbf{K}),\mathbf{x},Y\Big) = \prod_{\nu'}\Big(\sum_{\mathbf{k}} K_{\nu'\Gamma_1(\mathbf{k})}\Big)! \prod_{\mathbf{k}} \frac{1}{(K_{\nu'\Gamma_1(\mathbf{k})})!}\left(\frac{(k_\nu+1)x_{\nu'\Gamma_1(\mathbf{k}+\widehat{\mathbf{e}}_\nu)}}{\sum_{\mathbf{k}'} k'_\nu x_{\nu'\Gamma_1(\mathbf{k}')}}\right)^{K_{\nu'\Gamma_1(\mathbf{k})}} \tag{8e}$$

which complete the rules W for the first neighbourhood SIS model.

Figure 5 compares the results produced by this simplified model and the full one. Note that this is a stochastic version of the model presented in [11], except that the network structure is here static.

3.3 First neighbourhood SIR model

As in Sec. 3.2, we consider a full network model where the network structure is specified solely by the degree of its nodes. However, this time we consider SIR epidemiological dynamics: the accessible node states are $\nu \in \{S, I, R\}$, infection is the same as in SIS but recovery is replaced by removal (see the introduction of Sec. 3 for details).

We define the forward transition event of type $\mathcal{I}\mathbf{K}$ to be the $\mathcal{I}$nfection of the central node of a $S\Gamma_2(\mathbf{K})$ motif while a forward transition event of type $\mathcal{R}\mathbf{K}$ is the $\mathcal{R}$emoval of the central node of a $I\Gamma_2(\mathbf{K})$ motif. There is no backward

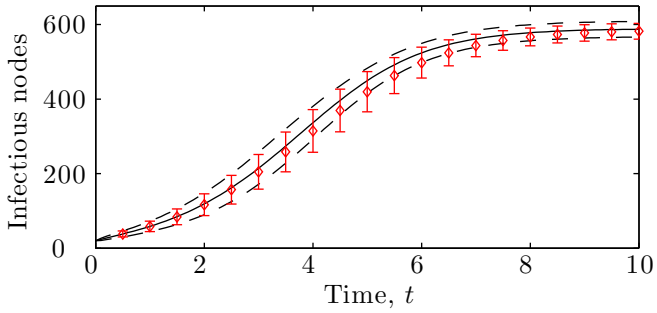


Figure 5: (Online version in colour.) Time evolution of the number of infectious nodes for SIS dynamics ($\alpha = 1$ and $\beta = 1$) on a CM network for which the number of nodes of each degree is prescribed by the sequence $\{0, 50, 200, 450, 300\}$ (total $N = 10^3$ nodes) with 2% of the nodes of each degree initially infectious. The mean and range of one standard deviation above and bellow the mean are shown. Curves: simplified system. Symbols: full system (10^5 simulations).

transition events. The model is specified by

$$\mathbf{x} = \left(\cdots, x_{S\Gamma_1(\mathbf{k})}, \cdots, x_{I\Gamma_1(\mathbf{k})}, \cdots, x_{R\Gamma_1(\mathbf{k})}, \cdots\right) \tag{9a}$$

$$\mathbf{r}^{\mathcal{I}\mathbf{K}} = \widehat{\mathbf{e}}_{I\widetilde{\Gamma}_1(\mathbf{K})} - \widehat{\mathbf{e}}_{S\widetilde{\Gamma}_1(\mathbf{K})} + \sum_{\nu}\sum_{\mathbf{k}} K_{\nu\Gamma_1(\mathbf{k})}\left(\widehat{\mathbf{e}}_{\nu\Gamma_1(\mathbf{k}+\widehat{\mathbf{e}}_I)} - \widehat{\mathbf{e}}_{\nu\Gamma_1(\mathbf{k}+\widehat{\mathbf{e}}_S)}\right) \tag{9b}$$

$$\mathbf{r}^{\mathcal{R}\mathbf{K}} = \widehat{\mathbf{e}}_{R\widetilde{\Gamma}_1(\mathbf{K})} - \widehat{\mathbf{e}}_{I\widetilde{\Gamma}_1(\mathbf{K})} + \sum_{\nu}\sum_{\mathbf{k}} K_{\nu\Gamma_1(\mathbf{k})}\left(\widehat{\mathbf{e}}_{\nu\Gamma_1(\mathbf{k}+\widehat{\mathbf{e}}_R)} - \widehat{\mathbf{e}}_{\nu\Gamma_1(\mathbf{k}+\widehat{\mathbf{e}}_I)}\right) \tag{9c}$$

$$q^{+}_{\mathcal{I}\mathbf{K}}(\mathbf{x}, Y) = \beta x_{S\widetilde{\Gamma}_1(\mathbf{K})}\Big(\sum_{\mathbf{k}} K_{I\Gamma_1(\mathbf{k})}\Big) P\Big(S\Gamma_2(\mathbf{K})\Big|\, S\widetilde{\Gamma}_1(\mathbf{K}), \mathbf{x}, Y\Big) \tag{9d}$$

$$q^{+}_{\mathcal{R}\mathbf{K}}(\mathbf{x}, Y) = \alpha\, x_{I\widetilde{\Gamma}_1(\mathbf{K})}\, P\Big(I\Gamma_2(\mathbf{K})\Big|\, I\widetilde{\Gamma}_1(\mathbf{K}), \mathbf{x}, Y\Big) \tag{9e}$$

$$q^{-}_{\mathcal{I}\mathbf{K}}(\mathbf{x}, Y) = q^{-}_{R\mathbf{K}}(\mathbf{x}, Y) = 0 \tag{9f}$$

where the inference terms are the same as in (8e).

3.4 First neighbourhood on-the-fly SIR model

We take a different perspective to the problem considered in Sec. 3.3 which requires to track much less elements in the state vector. Instead of considering "complete" first neighbourhood motifs, such as $\nu\Gamma_1(\mathbf{k})$, that specify the state of each of the central node's neighbours, we define the $\nu\Lambda_1(\kappa)$ motif as a central node of state ν for which we know that it has κ neighbours *unknown to us*. This last statement is important: were we to learn the state of one of these neighbours, this would cease to be a $\nu\Lambda_1(\kappa)$ motif and instead become a $\nu\Lambda_1(\kappa - 1)$ one. As usual, the state vector tracks the number of such motifs

$$\mathbf{x} = \left(\cdots, x_{S\Lambda_1(\kappa)}, \cdots, x_{I\Lambda_1(\kappa)}, \cdots, x_{R\Lambda_1(\kappa)}, \cdots\right). \tag{10a}$$

We recall from Sec. 3.2 how a CM network is built in a computer simulation: for each κ, n_κ nodes with κ stubs are created and the stubs are then randomly paired to form links. From this perspective, $\nu\Lambda_1(\kappa)$ may be reinterpreted as a ν state node with κ unpaired stubs: as stubs are removed once they are paired in the computer simulation, neighbours that were unknown are removed from these motifs once they become known to us. Hence,

$$\frac{\kappa x_{\nu\Lambda_1(\kappa)} - \delta_{\nu\nu'}\delta_{\kappa\kappa'}}{\sum_{\nu''}\sum_{\kappa''} \kappa'' x_{\nu''\Lambda_1(\kappa'')} - 1}$$

exactly gives the probability for an unknown neighbours of the central node of $\nu'\Lambda_1(\kappa')$ to be the central node of $\nu\Lambda_1(\kappa)$. Note that the Kronecker deltas $\left(\delta_{ii'} = \begin{smallmatrix} 1 & i=i' \\ 0 & i\neq i' \end{smallmatrix}\right)$ in the numerator and the -1 in the denominator both account for the stub of $\nu'\Lambda_1(\kappa')$ that we are pairing with a random stub.

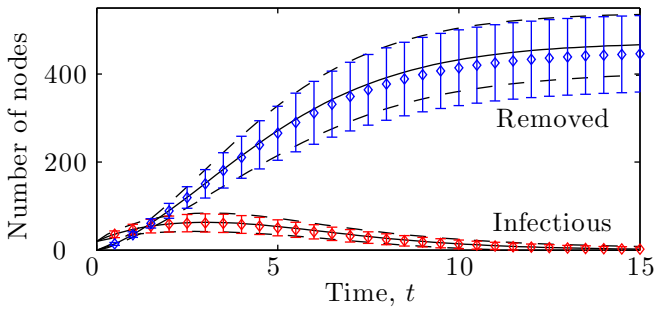


Figure 6: (Online version in colour.) Time evolution of the number of infectious and removed nodes for SIR dynamics ($\alpha = 1$ and $\beta = 1$) on a CM network for which the number of nodes of each degree is prescribed by the sequence $\{0, 50, 200, 450, 300\}$ (total $N = 10^3$ nodes) with 2% of the nodes of each degree initially infectious (all others are susceptible). The mean and range of one standard deviation above and bellow the mean are shown. Curves: simplified system. Symbols: full system (10^5 simulations).

A typical computer simulation would first build the network and then perform the SIR propagation dynamics on this network. However, we do not want to have to store the network structure for later consultation, which would require additional space in $\mathbf{x}$. Instead, we delay the network construction, leaving the stubs unpaired, and start the propagation dynamics right away. Just when the state of an unknown neighbour is required do we pair the corresponding stub with a randomly selected one, hence building the network *on-the-fly*. Since the knowledge of stubs being matched will be lost in the future, this information must only be required at the very moment it is obtained if we want the resulting dynamics to *exactly* reproduce the behaviour of the full system.

We thus take a different, although equivalent, perspective on the infection dynamics where each link is "probed" at most once. Instead of considering a probability $\beta\,dt$ of infection for each *susceptible* neighbours of infectious nodes, we consider the same probability for each of their *unknown* neighbours. Only when this probability returns true do we wonder about the state of the neighbour, whose state changes to infectious if and only if it was previously susceptible. In any case, we learned who were the neighbours of two nodes (*i.e.*, the infectious and its neighbour) and we must update the state vector accordingly.

Hence, we define the $\mathcal{I}$nfection transition event $\mathcal{I}\nu\kappa\kappa'$ such that an infectious at the center of a $I\Lambda_1(\kappa')$ motif attempts to infect the ν-state node at the center of a $\nu\Lambda_1(\kappa)$ motif. Of course, only $\mathcal{I}S\kappa\kappa'$ transition events result in real infections. The more traditional transition event $\mathcal{R}\kappa$ corresponds to the $\mathcal{R}$emoval of the infectious node at the center of a $I\Lambda_1(\kappa)$ motif, thus becoming $R\Lambda_1(\kappa)$. The model is specified by

$$\mathbf{r}^{\mathcal{I}S\kappa\kappa'} = \widehat{\mathbf{e}}_{I\Lambda_1(\kappa-1)} - \widehat{\mathbf{e}}_{S\Lambda_1(\kappa)} + \widehat{\mathbf{e}}_{I\Lambda_1(\kappa'-1)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa')} \tag{10b}$$

$$\mathbf{r}^{\mathcal{I}I\kappa\kappa'} = \widehat{\mathbf{e}}_{I\Lambda_1(\kappa-1)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa)} + \widehat{\mathbf{e}}_{I\Lambda_1(\kappa'-1)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa')} \tag{10c}$$

$$\mathbf{r}^{\mathcal{I}R\kappa\kappa'} = \widehat{\mathbf{e}}_{R\Lambda_1(\kappa-1)} - \widehat{\mathbf{e}}_{R\Lambda_1(\kappa)} + \widehat{\mathbf{e}}_{I\Lambda_1(\kappa'-1)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa')} \tag{10d}$$

$$\mathbf{r}^{\mathcal{R}\kappa} = \widehat{\mathbf{e}}_{R\Lambda_1(\kappa)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa)} \tag{10e}$$

$$q^{+}_{\mathcal{I}\nu\kappa\kappa'}(\mathbf{x}, Y) = \beta\,\kappa'\,x_{I\Lambda_1(\kappa')}\frac{\kappa x_{\nu\Lambda_1(\kappa)} - \delta_{I\nu}\delta_{\kappa\kappa'}}{\sum_{\nu''}\sum_{\kappa''}\kappa'' x_{\nu''\Lambda_1(\kappa'')} - 1} \tag{10f}$$

$$q^{+}_{\mathcal{R}\kappa}(\mathbf{x}, Y) = \alpha x_{I\Lambda_1(\kappa)} \tag{10g}$$

$$q^{-}_{\mathcal{I}\nu\kappa\kappa'}(\mathbf{x}, Y) = q^{-}_{\mathcal{R}\kappa}(\mathbf{x}, Y) = 0 \quad . \tag{10h}$$

The system (10) *exactly* reproduces the behaviour of the full system through the solution of (2). Since (3)–(4) are only approximations of (2), results obtained through these relationships are only approximative (Fig. 6 and Fig. 7). This model may be solved analytically for the mean value (see ESM [7] §III) and the results are in agreement with [19, 20]. Moreover, ESM [7] §IV shows how (10) may be rewritten with a state vector two thirds the size of (10a). This is a generalization to the case $\alpha \neq 0$ of the model presented in [14]. Further details are discussed in Section 4.4. We note that

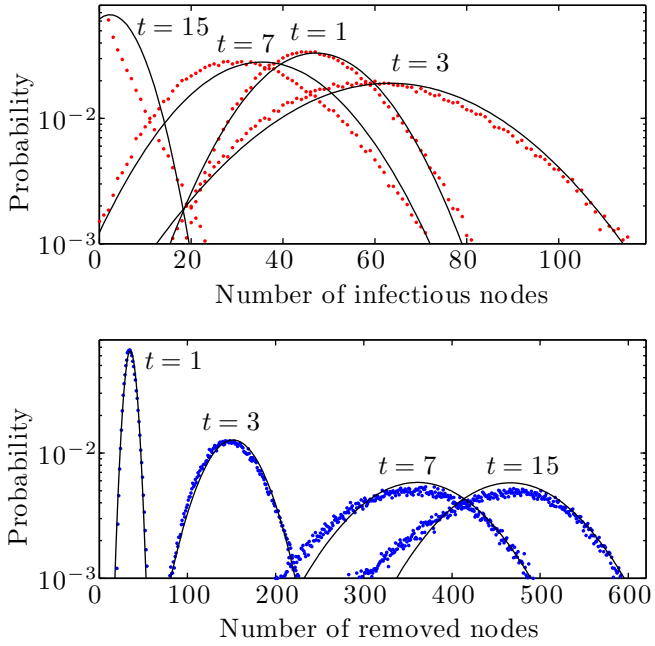


Figure 7: (Online version in colour.) Probability distribution at different times for the number of infectious and removed nodes. The parameters are the same as in Fig. 6. Curves: simplified system. Symbols: full system (10^5 simulations).

a conceptually similar approach has recently been developed independently [21] as a tool for a mathematically rigorous proof that a specific heterogeneous mean field model [19] holds in the limit of large network size.

4 Discussion

We now take a retrospective look at the results presented in Sec. 3 and obtain from these special examples general considerations concerning our modelling approach.

4.1 Accuracy of the results

One of the aims of this paper is to obtain simplified models that accurately reproduce the behaviour of complex systems. Since approximations are usually involved, it is to be expected that the results of the simplified model only agree with those of the full system over some range of parameters, where the approximations were valid.

The parameters used in Fig. 2–7 were chosen in order to investigate the limits of our approximations: while there is no perfect correspondence between the results of the full and simplified systems, their agreement is probably sufficient for both qualitative and quantitative applications. We distinguish between two categories of approximations: those inherent to the use of (3)–(4) and those due to the imperfect representation of Z through $\mathbf{x}$ and Y.

4.1.1 Gaussian approximation

Since (2) and (10) define a system that *exactly* reproduces the behaviour of the corresponding full system, any discrepancy in Fig. 6 must originate from the use of the Gaussian approximation (3)–(4). An important requirement for this approximation to be valid is that the size N of the system must be large.

Figures 2–7 all use networks of size $N = 1000$. As a rule of thumb, we found that (3)–(4) perform better for networks of at least a few hundred nodes, which is the case of many relevant real-world systems. Note that, for very small systems (tens of nodes), one could also directly and completely solve (2).

While a large network size N is required to justify treating the elements of $\mathbf{x}$ as real numbers, other phenomena may affect the validity of this approximation. For example, when the initial conditions are such that there is a single infectious

node, the continuous approximation fails at considering the probability for that node to recover (or to be removed) before transmitting the infection to one of its neighbours. Figures 2–7 circumvent this problem by using an initial condition with 20 infectious nodes: the probability for all of them to recover (or to be removed) before transmitting the infection is very low.

It is worth noting that the plateaux seen on Fig. 2 and 4–6 reflect different dynamical behaviours for the SIS and SIR systems. Indeed, while the total number of removed nodes reaches a maximum in the SIR system because there are no infectious left to recover, the steady state observed at the later times for our SIS models corresponds to a constant flow of recovery and new infections. In the former case, the approximation errors performed at earlier times accumulate. In the later, the exact path taken to attain equilibrium is of lesser importance and errors do not accumulate the same way.

4.1.2 Representation approximation

In general, the simplified system will not exactly reproduce the behaviour of the full system, even when using (2) instead of (3)–(4). This is the case of all our SIS models; while some of the discrepancy seen in Fig. 2–5 is explained by the Gaussian approximation, the imperfect representation of Z also contributes to the error.

Part of the problem can be understood as our failure to consider the correlation between the neighbours of a node and the time elapsed since this node has been in its present intrinsic state. For example, the neighbours of a susceptible node that has just recovered (*i.e.*, it was infectious a moment ago) may be much different than those of a susceptible node that has recovered a long time ago, while being similar to those of a node that is still infectious. Hence, one could hope to improve these SIS models through changes in Y alone (*i.e.*, with the same $\mathbf{x}$): first estimate the probability distribution for the time since when each node has last changed state and then infer the neighbourhoods accordingly. An alternative that could be simpler to implement, at the cost of increasing the size of $\mathbf{x}$, would consist in tracking more exhaustive motifs (*e.g.*, second neighbourhoods instead of first ones in Sec. 3.2).

However, there are more intricate consequences to the recovery of infectious nodes on a structure that is fixed in time: if at some point all the nodes of the same component (*i.e.*, a connected subnetwork that is disconnected from the rest of the network) are susceptible at the same time, then none of them may ever become infectious again. The connectivity of a network is strongly affected by the average degree of its nodes: our parameters correspond to an average degree of 5 for Fig. 2–4 (average degree of a neighbour also 5) and of 3 for Fig. 5 (average degree of a neighbour ≈ 3.23). When using smaller parameter values, this components-induced discrepancy becomes much larger since the simplified model then overestimates the number of infectious nodes. One could take the components into account by solving independent systems for each component (and merge the results afterwards) or by a clever adaptation of the inference process (see Sec. 4.6 for possible directions). Note that these effects are usually much less important when the network structure changes over time.

4.2 Pair-based models

Compared to the other models presented in Sec. 3, the two pair-based models of Sec. 3.1 use very small state vectors (*i.e.*, two or three elements). This is an important advantage of pair-based models in general: there are usually much less pair and node motifs than, *e.g.*, first neighbourhood ones, and tracking them thus requires much smaller $\mathbf{x}$.

Although we limited our study of pair-based models to regular and Erdős-Rényi networks, more complex network structures could also be considered. In the same way that (6) and (7) differ mostly by their inference terms, obtaining good inference from the little information stored in $\mathbf{x}$ is probably the principal challenge behind general and accurate pair-based stochastic models.

However, non-stochastic pair-based models are already possible on nontrivial network structures for SIR dynamics or, more generally, for processes such that a change in the state of one neighbour of a node can be treated as independent of that of another neighbour (SIS fails this assumption) [22]. Knowing (in Y) that a system behaves in this manner greatly simplifies the inference process, and this is the main reason for the success of the SIR pair-based model for the evolution of mean values on CM networks that is presented in [19, 20]. Whether or not the same approach may be used to obtain stochastic results is an open question.

4.3 First (and higher) neighbourhood models

By opposition, sufficiently accurate inference terms for first neighbourhood models are often straightforward to obtain. Although (8e) may be difficult to appreciate at first sight, it is the only inference term used in both Sec. 3.2 and Sec. 3.3. In fact, (8e) may well be the only inference term needed for generic first neighbourhood models for CM network structures.

Although first neighbourhood motifs are a "natural language" for expressing dynamics taking place on CM networks, they could also be used in the presence of other complex structures. This may be done through changes in $\mathbf{x}$ and/or Y; see ESM [7] §II for details.

The generality and ease of design of first-neighbourhood models comes at a cost: the state vector $\mathbf{x}$ is typically much larger than it would be in an equivalent pair-model. How large is $\mathbf{x}$ strongly depends on the maximal node degree present in the network and on the total number of accessible intrinsic node states (see ESM [7] §II for details). For typical values of these quantities, this does not cause major problems for the evaluation of the mean: numerically solving (3a) requires an acceptable amount of resources even for an $\mathbf{x}$ of dimension 10^6 and (3b) may often be simplified (*i.e.*, summed analytically).

However, evaluating the covariance matrix using (4) may cause problems: unless analytical simplifications are possible, solving this system scales as the square of the number of elements in $\mathbf{x}$. Future developments may decrease this bottleneck effect of the covariance matrix; see Sec. 4.5 for details. In any case, the size of $\mathbf{x}$ may be decreased by "coarse graining" the number of links between the central node and its neighbours; see ESM [7] §II G for details.

4.4 On-the-fly models

The on-the-fly model presented in Sec. 3.4 for SIR dynamics on CM networks *exactly* reproduces the behaviour of the full system. This is even more remarkable when one considers that the size of the state vector in the on-the-fly model is much smaller than in the alternative first neighbourhood model of Sec. 3.3. The reasons behind the success of the on-the-fly approach are similar to those discussed in Sec. 4.2 for the pair models presented in [22, 19, 20]: it is encoded in Y that, for each link, we *at most once* need to simultaneously know the state of the two nodes joined by that link [14].

The inference term (8e) is of "general purpose" in the sense that its Y does not provide information on the dynamical properties of the system, but only on how the motifs in $\mathbf{x}$ may be interconnected. This is why both (8) and (9) rely on (8e).

However, the inference terms of (10) have a specific character: Y contains information about (10) itself. Any change to the dynamics implies changes in the inference terms, with no guarantee that an acceptable solution exists. In fact, (10) was *designed* with this problem in mind. In other words, we obtained a simple and reliable model at the cost of "pre-computations" in the design process. Of all the possibilities in model-space, the information acquired by pointing at this specific one is what replaces the reduced size of the state vector. The same could be said of the deterministic SIR pair-based model on CM networks presented in [19, 20].

By contrast with the case discussed in Sec. 4.3, the small size of the state vector here allows for evaluations of the covariance matrix through (4), even when relatively high degree nodes are present. Alternatively, one may take advantage of the fact that, even for more complicated dynamics, the state vector of on-the-fly models can remain of manageable size for mean values calculations; see the introduction to ESM [7] §II for the concrete example of [13].

4.5 Complicated states *vs* complex assumptions

Section 4.4 revealed an unexpected depth to Y: one may achieve models of similar levels of accuracy by trading off complexity in the assumptions for a reduction in the size of the state vector $\mathbf{x}$. As an extreme example, if Y already gives the full behaviour of the system, then there is no need for tracking any information in $\mathbf{x}$. Without reaching such extremes, our on-the-fly model and the deterministic SIR pair-based model presented in [19, 20] both demonstrate the benefits of investing some time in the assumptions of our models.

While these examples required case-by-case analysis, one may benefit from the same realization in a general context: a first simplified model (W, $\mathbf{x}$ and Y) may generate the assumptions Y' to a different simplified model (W', $\mathbf{x}'$ and Y'). For example, when some dynamical process (*e.g.*, SIS or SIR) occurs on a network whose structure changes in time independently from this dynamics, one could obtain a first model for the structure alone and then feed the results to the second model, handling the remaining dynamics. Even more generally, one could compensate for the higher

computational requirements of (4) by first solving (3) on an elaborate model then feeding the resulting mean values to a simpler model for the sole purpose of estimating the covariance matrix.

4.6 Additional inference tools

While we introduced Y as a direct application of Bayes' rule, we have now seen that useful assumptions may be obtained by other means, including the solution of another system of the form (2). The next step in this direction would be to improve our inference process using alternative tools and models available to network science.

For example, branching processes [23] may be used to infer information concerning the connectivity and the components of the network structure. As discussed in Sec. 4.1.2, this point was a major shortfall of SIS models. This approach is even more interesting for the recently developed tools [24, 9] that are particularly compatible with the motifs and intrinsic node state approach presented in this paper.

Another tool of considerable interest are exponential random networks [25]. Indeed, these maximum entropy methods can simplify inferences that would have otherwise been prohibitively complex. Once again, this approach may be generalized to different kind of motifs and intrinsic node states.

5 Conclusion: general applicability

Although the examples of Sec. 3 focus on SIS and SIR dynamics, any specificity that could be modelled through a standard epidemiological compartmental model may *a priori* be considered by our approach: genders, age groups, vaccination, incubation period, disease phases, *etc*. Each compartment simply becomes an accessible node state in our formalism; see ESM [7] §I for details.

Furthermore, population dynamics considerations may be accounted for in a straightforward manner. Assuming first neighbourhood motifs, births and deaths of individuals correspond to events adding and removing motifs, respectively. Similarly, changes in interaction patterns amount to events replacing the affected motifs by new ones. In fact, from the model's perspective, there is no important distinction between a change in the interaction structure of the population and a change in the node states: both are events affecting motifs.

The generality of our systematic approach and the fact that its assumptions are explicitly stated suggests that it could be used as a common ground for comparing existing models too complex for direct comparison. Indeed, by considering such an existing model as the full system (specified by V_1 and Z_1), one may seek a simplified system (specified by W_1, X_1 and Y_1) approximately reproducing the original model (over a sufficient range of parameters).

If some transition event (in W_1) appears essential, this may reveal an important feature of the original model; the same holds true for motifs (in X_1) and prior knowledge (in Y_1). Moreover, assuming that this procedure has been done for a second existing model (specified by V_2 and Z_2), one may directly compare their simplified version in a common framework, which will help identify the assumptions required for their description. Note that this perspective is similar to a commutation diagram

$$\begin{array}{ccc} & \overset{\text{Difficult to compare}}{} & \\ (V_1, Z_1) & \longleftrightarrow\!\!\!\!\!\!/\;\; & (V_2, Z_2) \\ \text{Equivalent behaviour} \Updownarrow & & \Updownarrow \text{ Equivalent behaviour} \\ & \underset{\text{Comparable}}{} & \\ (W_1, X_1, Y_1) & \longleftrightarrow & (W_2, X_2, Y_2) \end{array} .$$

For example, if $X_1 = X_2$ and $Y_1 = Y_2$, we know that the discrepancies between the two original models is imputable to the difference in the transition events. Finding a minimal set of changes to W_1 and/or W_2 causing both models to agree may then help identify the very cause of the discrepancies.

Acknowledgements

The research team is grateful to the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Fonds de recherche du Québec — Nature et technologies (FRQ—NT) for financial support.

References

[1] Broeck W, Gioannini C, Goncalves B, Quaggiotto M, Colizza V, Vespignani A. The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BMC Infectious Diseases. 2011;11(1):37.

[2] Riley S. Large-Scale Spatial-Transmission Models of Infectious Disease. Science. 2007;316(5829):1298–1301.

[3] Lee BY, Brown ST, Cooley P, Grefenstette JJ, Zimmerman RK, Zimmer SM, et al. Vaccination Deep Into a Pandemic Wave: Potential Mechanisms for a "Third Wave" and the Impact of Vaccination. American Journal of Preventive Medicine. 2010;39(5):e21 – e29.

[4] Barrat A, Barthélemy M, Vespignani A. Dynamical Processes on Complex Networks. New York: Cambridge University Press; 2008.

[5] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: Structure and dynamics. Phys Rep. 2006;424:175–308.

[6] Gardiner CW. Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences. Springer-Verlag; 2004.

[7] Noël PA, Allard A, Hébert-Dufresne L, Vincent M, Dubé LJ. Electronic Supplementary Material to "Epidemics on contact networks: a general stochastic approach". 2011;URL to be inserted by editors.

[8] House T, Davies G, Danon L, Keeling MJ. A Motif-Based Approach to Network Epidemics. Bull Math Biol. 2009;71:1693–1706.

[9] Karrer B, Newman MEJ. Random graphs containing arbitrary distributions of subgraphs. Phys Rev E. 2010;82(6):066118.

[10] Gleeson JP. High-Accuracy Approximation of Binary-State Dynamics on Networks. Phys Rev Lett. 2011;107:068701.

[11] Marceau V, Noël PA, Hébert-Dufresne L, Allard A, Dubé LJ. Adaptive networks: Coevolution of disease and topology. Phys Rev E. 2010;82(3):036116.

[12] Hébert-Dufresne L, Noël PA, Marceau V, Allard A, Dubé LJ. Propagation dynamics on networks featuring complex topologies. Phys Rev E. 2010;82(3):036115.

[13] Marceau V, Noël PA, Hébert-Dufresne L, Allard A, Dubé LJ. Modeling the dynamical interaction between epidemics on overlay networks. Phys Rev E. 2011;84(2):026105.

[14] Noël PA, Allard A, Hébert-Dufresne L, Vincent M, Dubé LJ. Propagation on networks: an exact alternative perspective. e-print arXiv:11020987. 2011;Submitted for publication.

[15] Keeling MJ, Eames KTD. Networks and epidemic models. J R Soc Interface. 2005;2(4):295–307.

[16] Bansal S, Grenfell BT, Meyers LA. When individual behaviour matters: homogeneous and network models in epidemiology. J R Soc Interface. 2007;4:879–891.

[17] Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, et al. Networks and the Epidemiology of Infectious Disease. Interdiscip Perspect Infect Dis. 2011;2011:284909–1–28.

[18] Dangerfield CE, Ross JV, Keeling MJ. Integrating stochasticity and network structure in an epidemic model. J R Soc Interface. 2009;6:761–774.

[19] Volz E. SIR dynamics in random networks with heterogeneous connectivity. J Math Biol. 2008;56(3):293–310.

[20] Miller JC. A note on a paper by Erik Volz: SIR dynamics in random networks. J Math Biol. 2010;62(3):349–358.

[21] Decreusefond L, Dhersin J, Moyal P, Tran CV. Large graph limit for an SIR process in random network with heterogeneous connectivity. Ann Appl Probab;Accepted for publication.

[22] Miller JC, Slim AC, Volz EM. Edge-based compartmental modeling for infectious disease spread. J R Soc Interface;Doi:10.1098/rsif.2011.0403.

[23] Newman MEJ, Strogatz SH, Watts DJ. Random graphs with arbitrary degree distributions and their applications. Phys Rev E. 2001;64:026118.

[24] Allard A, Noël PA, Dubé LJ, Pourbohloul B. Heterogeneous bond percolation on multitype networks with an application to epidemic dynamics. Phys Rev E. 2009;79:036113.

[25] Park J, Newman MEJ. Statistical mechanics of networks. Phys Rev E. 2004;70:066117(1–13).

Electronic Supplementary Material to "Epidemics on contact networks: a general stochastic approach"

Pierre-André Noël, Antoine Allard, Laurent Hébert-Dufresne, Vincent Marceau and Louis J. Dubé
Département de Physique, de Génie Physique et d'Optique, Université Laval, Québec (QC), Canada

December 8, 2011

Abstract

This document provides supplementary informations to "Markov processes on complex networks with applications to spreading dynamics" [1], which will hereafter be referred to as the "main text". Equation numbers from the main text are preceded by the letters "MT".

1 State of the full system (Z)

The state of the full system is specified by the intrinsic state of all of its components and the network structure governing their interactions. As defined in the main text [1] §II A, a node has at any time a single intrinsic node state ν. We note $\mathcal{N}$ the total number of accessible node states, which takes its minimal value $\mathcal{N} = 1$ when nodes are intrinsically indiscernible. In addition, we introduce the *intrinsic link state* ℓ. In a similar way, $\mathcal{L}$ is the total number of accessible link states, which takes its minimal value $\mathcal{L} = 1$ for intrinsically indiscernible links ($\mathcal{L} = 1$ throughout the main text). Examples of link states may include any relevant characteristics of the interactions: their context (*e.g.*, professional, friendship, partnership, ...), their dynamical status (*e.g.*, active or inactive), their weight (*e.g.*, strong, normal, weak, ...), *etc.*

While each node (or link) may be in only one intrinsic state at any given time, different pieces of information may be encoded in this single state. As it is the case in standard compartmental models, a simple Cartesian product may suffice to this task. For example, an SIR dynamics (three epidemiological states, see main text [1] §III) where we discern two genders (male and female) would result in a total of $\mathcal{N} = 6$ accessible node states. Note that there is no particular problem caused by the fact that one of these characteristics is susceptible to change during the process while the other one remains constant.

In addition to intrinsic link states, links may (or may not) be directed. The links of the networks discussed in the main text [1] are *undirected*: the interaction between the two linked nodes is bidirectional and symmetric. *Directed* links represent interactions that are either unidirectional or asymmetrical. A network that has only undirected links is said to be *undirected*, one that has only directed links is *directed* and one that has both is *semi-directed*.

2 More on motifs

We introduce new motifs and generalize those presented in the main text [1] for intrinsic link types and directed (or semi-directed) networks. Table 1 summarizes the total number of motifs in each of these classes.

Of high practical importance is the fact that the entries of this table differ widely in their scaling behaviours. For example, the on-the-fly model [2] for two interacting SIR dynamics each propagating on their own network structure uses $\mathcal{N} = 9$ (two SIR), $\mathcal{L} = 3$ (links of the first network alone, links of the second network alone and overlapping links) and $\mathcal{D} = 1$ (undirected network). Looking up in table 1, we see that this requires of the order of $\mathcal{K}^3$ on-the-fly (degreed node) motifs, where $\mathcal{K}$ denotes the highest accessible node degree. By opposition, implementing a first neighbourhood version of this model would require of the order of $\mathcal{K}^{27}$ motifs!

Table 1: Number of motifs in selected classes. We defined $\mathcal{D} = 1$ for undirected, $\mathcal{D} = 2$ for directed and $\mathcal{D} = 3$ for semi-directed networks. Moreover, $\mathcal{K}$ denotes the highest accessible node degree.

Motifs	Number
Node	$\mathcal{N}$
Pair	$\frac{1}{2}\mathcal{L}\mathcal{N}\big(\mathcal{D}\mathcal{N} + (\mathcal{D}-2)^2\big)$
Triple	$\frac{1}{2}\mathcal{D}\mathcal{L}\mathcal{N}^2(\mathcal{D}\mathcal{L}\mathcal{N}+1)$
Triangle	$\frac{1}{3}\mathcal{D}\mathcal{L}\mathcal{N}+\frac{1}{2}\mathcal{D}\big((\mathcal{D}-2)\mathcal{L}\mathcal{N}\big)^2+\frac{1}{6}(\mathcal{D}\mathcal{L}\mathcal{N})^3$
Degreed node	$\mathcal{N}\binom{\mathcal{K}+\mathcal{D}\mathcal{L}}{\mathcal{K}}$
Degreed pair	$\frac{\mathcal{L}\mathcal{N}}{2}\binom{\mathcal{K}-1+\mathcal{D}\mathcal{L}}{\mathcal{K}-1}\Big(\mathcal{D}\mathcal{N}\binom{\mathcal{K}-1+\mathcal{D}\mathcal{L}}{\mathcal{K}-1}+(\mathcal{D}-2)^2\Big)$
First neighbourhood (node)	$\mathcal{N}\binom{\mathcal{K}+\mathcal{D}\mathcal{L}\mathcal{N}}{\mathcal{K}}$
First neighbourhood pair	$\frac{\mathcal{L}\mathcal{N}}{2}\binom{\mathcal{K}-1+\mathcal{D}\mathcal{L}\mathcal{N}}{\mathcal{K}-1}\Big(\mathcal{D}\mathcal{N}\binom{\mathcal{K}-1+\mathcal{D}\mathcal{L}\mathcal{N}}{\mathcal{K}-1}+(\mathcal{D}-2)^2\Big)$
Second neighbourhood (node)	$\mathcal{N}\binom{\mathcal{K}+\mathcal{D}\mathcal{L}\mathcal{N}\binom{\mathcal{K}-1+\mathcal{D}\mathcal{L}\mathcal{N}}{\mathcal{K}-1}}{\mathcal{K}}$

2.1 Pair motifs

We note $\nu \overset{\ell}{-} \nu'$ (resp. $\nu \overset{\ell}{\to} \nu'$) an undirected (resp. directed) pair motif formed of a state ν node linked through a state ℓ link to a state ν' node. In the case of directed motifs, the direction of the arrow usually specifies the "strongest causal effect" of this asymmetric interaction (although this needs not be the case). All the undirected and directed motifs are possible in a semi-directed network. We may omit the index ℓ over the links when $\mathcal{L} = 1$.

2.2 Triple motifs

We note $\nu \overset{\ell}{-} \nu' \overset{\ell'}{-} \nu''$ an undirected triple motif formed of a state ν node linked through a state ℓ link to a state ν' node itself linked through a state ℓ' link to a state ν'' node. As for pair motifs, each of these nodes may have other neighbours than those that are explicitly specified. The notation directly generalizes to directed (*e.g.*, $\nu \overset{\ell}{\to} \nu' \overset{\ell'}{\leftarrow} \nu''$) and semi-directed (*e.g.*, $\nu \overset{\ell}{-} \nu' \overset{\ell'}{\to} \nu''$) triple motifs.

The term "2-star" is often used to refer to a triple motif for which both extremities (*e.g.*, the nodes of state ν and ν'' in the motif $\nu \overset{\ell}{-} \nu' \overset{\ell'}{-} \nu''$) are explicitly *forbidden* to be neighbours. In models that also use triangle motifs, 2-star motifs may explicit the absence of the last link that would form a triangle. Another common use of triple motifs comes in the inference process of (usually deterministic) pair based models.

2.3 Triangle motifs and other small subnetworks

Three nodes that are all neighbours of each other form a triangle motif. An horizontal bracket represents the additional link that would be missing in the analogous triple motif, *e.g.*,

$$\underbrace{\nu \overset{\ell}{-} \nu' \overset{\ell'}{-} \nu''}_{\ell''}$$

for an undirected network.

Triangle motifs are usually considered in models that should account for clustering. Their number may either directly be tracked in the state vector $\mathbf{x}$ [3] or their implicit presence (stated in Y, *e.g.*, through a clustering coefficient) may be accounted for in the inference process [4].

The same notation may be generalized to other motifs consisting of small subnetworks, *e.g.*, square motifs [3].

2.4 Clique motifs

A vague definition of a clique motif is "a subgroup of nodes that share more links among themselves than what could be expected otherwise for the same number of randomly selected nodes". In applications, one may refine this definition according to the specificities of the problem at hand, *e.g.*, "an Erdős-Rényi subnetwork (link probability p) of n_S susceptible nodes and n_I infectious nodes". Clique motifs are usually considered in models that should account for community structure [5].

2.5 Degreed motifs

A degreed motif is a motif for which we know the degree of all the nodes forming the motif: a degreed node motif is a node of specified state and degree; a degreed link motif is two nodes of specified state and degree that are known to be neighbours; *etc*. The in-degree and out-degree are both specified in directed and semi-directed networks; the latter cases also specify undirected degrees. Likewise, degrees pertaining to different types of links are specified independently.

In the same way that pair motifs are usually combined with node motifs, degreed pair motifs are usually combined with degreed node motifs [6]. Note that the on-the-fly motifs $\nu\Lambda_1(\kappa)$ presented in the main text [1] §III D can be understood as a special case of degreed node motifs where the degree is replaced by the "degree to unknown nodes".

2.6 n-th neighbourhood motifs

Similarly to degreed motifs, an n-th neighbourhood motif is a motif for which we specify the state of all the n-th neighbours of the nodes forming the original motif. Hence, the notation $\underline{\nu}\Gamma_1(\mathbf{k})$ $\big[$resp. $\underline{\nu}\Gamma_2(\mathbf{K})\big]$ of the main text corresponds to a first (resp. second) neighbourhood *node* motif. These concepts are directly generalizable to types of links and to directed or semi-directed networks.

First neighbourhood node motifs can be understood as tracking the correlation between the state of a node and the state of all its neighbours. By opposition, degreed pair motifs track the correlation between the state and degree of two neighbouring nodes. While similar information may be obtained from both motif classes, a model based on one may perform better than a model based on the other depending on the characteristics of the full system to be modelled.

2.7 Coarse-grained degree and/or neighbourhood

Not all entries of table 1 depend on the maximal degree $\mathcal{K}$, but those that do quickly increase for large $\mathcal{K}$. This is problematic since many real-world systems contain high degree nodes.

However, one may overcome this limitation by coarse-graining degrees through ranges: the range containing the degree of a node is specified instead of the degree itself. For example, given the ranges

$$\underbrace{[0,0]}_{\text{range 0}},\underbrace{[1,1]}_{\text{range 1}},\underbrace{[2,3]}_{\text{range 2}},\underbrace{[4,7]}_{\text{range 3}},\underbrace{[8,15]}_{\text{range 4}},\underbrace{[16,31]}_{\text{range 5}}\text{ and }\underbrace{[32,63]}_{\text{range 6}}\quad,$$

we would say of a degree 23 node that its degree lies within range 5. Hence, for the purpose of evaluating the number of motifs in table 1, one should here use $\mathcal{K}=6$ (one less than the total number of ranges) instead of $\mathcal{K}=63$ (highest representable degree).

While the previous example used powers of 2 for simplicity, a slower increase is probably desirable in most applications. However, since the number of neighbourhood and degreed motifs strongly depends on $\mathcal{K}$, even the slightest reduction in this number may be significant. Note that this coarse-graining method is of particular interest when the real-world data used to calibrate the model is already coarse-grained, which is commonly the case for census data.

3 Deterministic solution of on-the-fly SIR model

This section provides the deterministic solution of (MT10) from the main text [1]. Our result is the same one as for the SIR pair model presented in [7, 8].

We first rewrite (MT3) for the specific case of (MT10)

$$\frac{d\boldsymbol{\mu}(t)}{dt} = \sum_{\nu}\sum_{\kappa}\sum_{\kappa'} \mathbf{r}^{\mathcal{I}\nu\kappa\kappa'} q^{+}_{\mathcal{I}\nu\kappa\kappa'}\big(\boldsymbol{\mu}(t), Y\big) + \sum_{\kappa} \mathbf{r}^{\mathcal{R}\kappa} q^{+}_{\mathcal{R}\kappa}\big(\boldsymbol{\mu}(t), Y\big) \tag{1}$$

then collect the contributions to $\mu_{S\Lambda_1(\kappa)}$, $\mu_{I\Lambda_1(\kappa)}$ and $\mu_{R\Lambda_1(\kappa)}$ (dropping the functional dependencies for brevity)

$$\frac{d\mu_{S\Lambda_1(\kappa)}}{dt} = -\sum_{\kappa'} q^{+}_{\mathcal{I}S\kappa\kappa'} \tag{2a}$$

$$\frac{d\mu_{I\Lambda_1(\kappa)}}{dt} = \sum_{\kappa'}\Big(q^{+}_{\mathcal{I}S(\kappa+1)\kappa'} + q^{+}_{\mathcal{I}I(\kappa+1)\kappa'} - q^{+}_{\mathcal{I}I\kappa\kappa'}\Big) + \sum_{\nu}\sum_{\kappa'}\Big(q^{+}_{\mathcal{I}\nu\kappa'(\kappa+1)} - q^{+}_{\mathcal{I}I\kappa'\kappa}\Big) - q^{+}_{\mathcal{R}\kappa} \tag{2b}$$

$$\frac{d\mu_{R\Lambda_1(\kappa)}}{dt} = \sum_{\kappa'}\Big(q^{+}_{\mathcal{I}R(\kappa+1)\kappa'} - q^{+}_{\mathcal{I}R\kappa\kappa'}\Big) + q^{+}_{\mathcal{R}\kappa} \quad . \tag{2c}$$

Using the definitions

$$\lambda = \sum_{\kappa} \kappa\mu_{I\Lambda_1(\kappa)} \quad \text{and} \quad \omega = \sum_{\nu}\sum_{\kappa} \kappa\mu_{\nu\Lambda_1(\kappa)} \tag{3}$$

where λ is the total number of stubs belonging to infectious nodes and ω is the total number of stubs in the system, (2) becomes

$$\frac{d\mu_{S\Lambda_1(\kappa)}}{dt} = -\frac{\beta\lambda\kappa\mu_{S\Lambda_1(\kappa)}}{\omega} \tag{4a}$$

$$\frac{d\mu_{I\Lambda_1(\kappa)}}{dt} = \frac{\beta\lambda(\kappa+1)\mu_{S\Lambda_1(\kappa+1)}}{\omega} - \alpha\mu_{I\Lambda_1(\kappa)} + \beta\left(1+\frac{\lambda}{\omega}\right)\Big((\kappa+1)\mu_{I\Lambda_1(\kappa+1)} - \kappa\mu_{I\Lambda_1(\kappa)}\Big) \tag{4b}$$

$$\frac{d\mu_{R\Lambda_1(\kappa)}}{dt} = \frac{\beta\lambda}{\omega}\Big((\kappa+1)\mu_{R\Lambda_1(\kappa+1)} - \kappa\mu_{R\Lambda_1(\kappa)}\Big) + \alpha\mu_{I\Lambda_1(\kappa)} \quad . \tag{4c}$$

Note that (MT10f) has been approximated by dropping the Kronecker delta in the numerator and the -1 in the denominator.

We now consider the evolution of the total number of stubs ω by summing the contributions from (4)

$$\frac{d\omega}{dt} = \sum_{\nu}\sum_{\kappa} \kappa\frac{d\mu_{\nu\Lambda_1(\kappa)}}{dt} = -2\beta\lambda \quad . \tag{5}$$

One may understand (5) as "during the time interval $[t, t+dt)$, each one of the λ stubs belonging to infectious nodes have probability βdt to be paired to another stub, thus causing a decrease by 2 of ω. Noting $\omega_0 = \omega(0)$ the total number of stubs in the initial condition, we introduce the change of variable

$$\theta = \sqrt{\frac{\omega}{\omega_0}} \quad \text{such that} \quad \frac{d\theta}{dt} = -\frac{\beta\lambda}{\theta\omega_0} \tag{6}$$

Notice that $t = 0$ corresponds to $\theta = 1$ and that θ *decreases* with time. Using this change of variable in (4a) gives

$$\frac{d\mu_{S\Lambda_1(\kappa)}}{d\theta} = \frac{\kappa\mu_{S\Lambda_1(\kappa)}}{\theta} \tag{7}$$

which has the solution

$$\mu_{S\Lambda_1(\kappa)}(t) = x_{S\Lambda_1(\kappa)}(0)\big(\theta(t)\big)^{\kappa} \quad . \tag{8}$$

using the initial condition $\mu_{S\Lambda_1(\kappa)}(0) = x_{S\Lambda_1(\kappa)}(0)$.

For convenience, we define

$$f(\theta) = \sum_{\kappa} x_{S\Lambda_1(\kappa)}(0)\theta^{\kappa} \quad . \tag{9}$$

Noticing that

$$\sum_\kappa \kappa(\kappa+1)\mu_{S\Lambda_1(\kappa+1)} = \theta^2 \sum_\kappa (\kappa-1)\kappa x_{S\Lambda_1(\kappa)}(0)\theta^{\kappa-2} = \theta^2 f''(\theta) \quad , \tag{10}$$

we obtain the evolution of the total number λ of stubs belonging to infectious nodes by summing the contributions (4b)

$$\frac{d\lambda}{dt} = \sum_\kappa \kappa \frac{d\mu_{I\Lambda_1(\kappa)}}{dt} = \frac{\beta\lambda\theta^2 f''(\theta)}{\omega} - \beta\lambda\left(1+\frac{\lambda}{\omega}\right) - \alpha\lambda \quad . \tag{11}$$

Again using the change of variable (6), we get

$$\frac{d\lambda}{d\theta} = \frac{\lambda}{\theta} + \theta\omega_0\left(1+\frac{\alpha}{\beta}\right) - \theta f''(\theta) \tag{12}$$

which has the solution

$$\lambda = \theta^2\omega_0\left(1+\frac{\alpha}{\beta}\right) - \theta\omega_0\frac{\alpha}{\beta} - \theta f'(\theta) \tag{13}$$

for an initial condition without removed nodes $\big[$*i.e.*, $\lambda(0) = \omega_0 - f'(1)\big]$.

Using this solution in (6) gives

$$\frac{d\theta}{dt} = -\beta\theta + \alpha(1-\theta) + \beta\frac{f'(\theta)}{\omega_0} \tag{14}$$

whose solution provides $\theta(t)$. Using

$$S(t) = f\big(\theta(t)\big) \tag{15a}$$

$$I(t) = N - S(t) - R(t) \tag{15b}$$

$$\frac{dR(t)}{dt} = \alpha I(t) \quad , \tag{15c}$$

we finally obtain the total number $S = \sum_\kappa \mu_{S\Lambda_1(\kappa)}$ of susceptible, $I = \sum_\kappa \mu_{I\Lambda_1(\kappa)}$ of infectious and $R = \sum_\kappa \mu_{R\Lambda_1(\kappa)}$ of removed nodes at any given time t. A direct application of (8)–(9) provides (15a), conservation of the nodes provides (15b) and using the definitions of $I(t)$ and $R(t)$ in (4c) provides (15c). Although obtained differently, this solution corresponds to that of the pair-based SIR model presented in [7, 8].

4 Alternative form of on-the-fly SIR model

The $R\Gamma_1(\mathbf{k})$ motifs were included in (MT10a) for the sake of clarity alone: noting s an unmatched stub motif, we could rewrite (MT10) with a state vector two thirds the size of (MT10a)

$$\begin{aligned}
\mathbf{x} &= \big(\cdots, x_{S\Lambda_1(\kappa)}, \cdots, x_{I\Lambda_1(\kappa)}, \cdots, x_s\big) \\
\mathbf{r}^{\mathcal{I}S\kappa\kappa'} &= \widehat{\mathbf{e}}_{I\Lambda_1(\kappa-1)} - \widehat{\mathbf{e}}_{S\Lambda_1(\kappa)} + \widehat{\mathbf{e}}_{I\Lambda_1(\kappa'-1)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa')} - 2\widehat{\mathbf{e}}_s \\
\mathbf{r}^{\mathcal{I}I\kappa\kappa'} &= \widehat{\mathbf{e}}_{I\Lambda_1(\kappa-1)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa)} + \widehat{\mathbf{e}}_{I\Lambda_1(\kappa'-1)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa')} - 2\widehat{\mathbf{e}}_s \\
\mathbf{r}^{\mathcal{I}R\kappa\kappa'} &= \widehat{\mathbf{e}}_{I\Lambda_1(\kappa'-1)} - \widehat{\mathbf{e}}_{I\Lambda_1(\kappa')} - 2\widehat{\mathbf{e}}_s \\
\mathbf{r}^{\mathcal{R}\kappa} &= -\widehat{\mathbf{e}}_{I\Lambda_1(\kappa)} \\
q^+_{\mathcal{I}S\kappa\kappa'}(\mathbf{x}, Y) &= \beta\,\kappa'\,x_{I\Lambda_1(\kappa')}\frac{\kappa x_{S\Lambda_1(\kappa)}}{x_s - 1} \\
q^+_{\mathcal{I}I\kappa\kappa'}(\mathbf{x}, Y) &= \beta\,\kappa'\,x_{I\Lambda_1(\kappa')}\frac{\kappa x_{I\Lambda_1(\kappa)} - \delta_{\kappa\kappa'}}{x_s - 1} \\
q^+_{\mathcal{I}R\kappa\kappa'}(\mathbf{x}, Y) &= \beta\,\kappa'\,x_{I\Lambda_1(\kappa')}\frac{x_s - 1 - \sum_{\kappa''}\kappa''\left(x_{S\Lambda_1(\kappa'')} + x_{I\Lambda_1(\kappa'')}\right)}{x_s - 1} \\
q^+_{\mathcal{R}\kappa}(\mathbf{x}, Y) &= \alpha x_{I\Lambda_1(\kappa)} \\
q^-_{\mathcal{I}\nu\kappa\kappa'}(\mathbf{x}, Y) &= q^-_{\mathcal{R}\kappa}(\mathbf{x}, Y) = 0 \quad .
\end{aligned}$$

Note that x_s plays the same role as ω in Sec. 3. Proceeding similarly, the state vector $\mathbf{x} = \big(\cdots, x_{S\Lambda_1(\kappa)}, \cdots, x_s\big)$ would suffice for a SI model (*i.e.*, $\alpha = 0$) [9].

References

[1] Noël PA, Allard A, Hébert-Dufresne L, Vincent M, Dubé LJ. Markov processes on complex networks with applications to spreading dynamics. Exact reference to be inserted by editors. 2011;.

[2] Marceau V, Noël PA, Hébert-Dufresne L, Allard A, Dubé LJ. Modeling the dynamical interaction between epidemics on overlay networks. Phys Rev E. 2011;84(2):026105.

[3] House T, Davies G, Danon L, Keeling MJ. A Motif-Based Approach to Network Epidemics. Bull Math Biol. 2009;71:1693–1706.

[4] Keeling MJ, Rand DA, Morris AJ. Correlation models for childhood epidemics. Proc R Soc B. 1997;264(1385):1149–1156.

[5] Hébert-Dufresne L, Noël PA, Marceau V, Allard A, Dubé LJ. Propagation dynamics on networks featuring complex topologies. Phys Rev E. 2010;82(3):036115.

[6] Eames KTD, Keeling MJ. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. PNAS. 2002;99:13330–13335.

[7] Volz E. SIR dynamics in random networks with heterogeneous connectivity. J Math Biol. 2008;56(3):293–310.

[8] Miller JC. A note on a paper by Erik Volz: SIR dynamics in random networks. J Math Biol. 2010;62(3):349–358.

[9] Noël PA, Allard A, Hébert-Dufresne L, Vincent M, Dubé LJ. Propagation on networks: an exact alternative perspective. e-print arXiv:11020987. 2011;Submitted for publication.